
Proceedings Paper:
Sutton, Selina, Foulkes, Paul orcid.org/0000-0001-9481-1004, Kirk, David et al. (1 more author) (2019) Voice as a design material : sociophonetic inspired design strategies in Human-Computer Interaction. In: CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019). CHI Conference on Human Factors in Computing Systems Proceedings . ACM .

https://doi.org/10.1145/3290605.3300833

Voice as a Design Material: Sociophonetic Inspired Design Strategies in Human-Computer Interaction

Selina Jeanne Sutton
Northumbria University
Newcastle upon Tyne, United Kingdom
selina.sutton@northumbria.ac.uk

Paul Foulkes
University of York
York, United Kingdom
paul.foulkes@york.ac.uk

David Kirk
Northumbria University
Newcastle upon Tyne, United Kingdom
david.kirk@northumbria.ac.uk

Shaun Lawson
Northumbria University
Newcastle upon Tyne, United Kingdom
shaun.lawson@northumbria.ac.uk

ABSTRACT

While there is a renewed interest in voice user interfaces (VUI) in HCI, little attention has been paid to the design of VUI voice output beyond intelligibility and naturalness. We draw on the field of *sociophonetics* - the study of the social factors that influence the production and perception of speech - to highlight how current VUIs are based on a limited and homogenised set of voice outputs. We argue that current systems do not adequately consider the diversity of peoples' speech, how that diversity represents sociocultural identities, and how voices have the potential to shape user perceptions and experiences. Ultimately, as other technological developments have influenced the ideologies of language, the voice outputs of VUIs will influence the ideologies of speech. Based on our argument, we pose three design strategies for VUI voice output design - *individualisation*, *context awareness*, and *diversification* - to motivate new ways of conceptualising and designing these technologies.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**; *Sound-based input / output*.

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300833

KEYWORDS

Voice user interface; sociophonetics; design material; experience-centered design; computer synthesized speech.

ACM Reference Format:
Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. 2019. Voice as a Design Material: Sociophonetic Inspired Design Strategies in Human-Computer Interaction. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3290605.3300833

1 INTRODUCTION

A voice user interface (VUI) allows human interaction with computers via the medium of voice. The commercial availability of VUI devices has dramatically increased in recent years with many technology companies producing their own voice-based assistants. These are available on smart phones (e.g. Siri [4], Google Assistant [28], Cortana [47]), smart speakers (e.g. Amazon Alexa [3]), and there may be integration across assistants in the future (e.g. [91]). The global market for these assistants is predicted to reach $4.61 billion by the early 2020s [18]. Both the technical and user-centred challenges addressed prior to the wide commercial availability of these devices evidences the complexity of designing and developing a VUI. These include recognising speech input [94], navigating the interaction [74], and producing intelligible voice output [33]. Indeed, there has been a growing interest in voice-based interaction at CHI (e.g. papers [61, 76, 82], courses [57–60], workshops [54, 56], and panels [55]). However, evaluation methods have not evolved in line with the advances made in voice output, and so less attention has been paid to properties of voice beyond intelligibility and naturalness [2]. While great progress has been made in their usability and reliability, in this paper we argue other aspects of VUI design - namely the voice output - should be considered more critically.

The need to pay attention to the design of voice has been highlighted previously [14, 29, 62, 63]. But thus far, there has been little guidance of how to do this [12] and what the repercussions of VUI voices might be. We address this gap by drawing on knowledge and insights from the field of *Sociophonetics* - the study of the social factors that influence the production and perception of speech [27]. In this paper, we argue how sociophonetics provides a range of conceptual and methodological tools which are valuable for critically exploring the design of voice output. How voices are perceived by people can be highly dependent on their own personal histories, and the social and cultural contexts where a voice is heard. Voices can be grounded in stereotypes, prejudices, and speech ideologies, and can give emphasis to certain social groups and cultural identities over others. Based on this, we argue sociophonetics can be integrated into the process of voice design to different degrees, and we provide examples of what this might look like through three design strategies.

Our contributions to the emerging literature on VUI design are two-fold. First, we offer the first attempt at bringing into dialogue the growing literature on VUI design in HCI and the field of sociophonetics, two fields that will become increasingly entwined as more artificial voices need to be designed and these voices affect and influence society. Second, we build on the growing critical discourse around speech technology in HCI (e.g. [5]) by offering a theoretically underpinned approach to considering the design of voice outputs and their wider sociocultural implications.

2 AN OVERVIEW OF THIS PAPER

To avoid confusion, it's useful for us to outline what this paper is *not* about. First, our focus is on speech, not language. Language is the message that we want to communicate, via words and sentences, whereas speech [1] comprises the *sounds* (i.e. the series of acoustic events) that emerge from mouths to transmit those words and sentences [10]. Second, our aim in this paper is not to address the argument that VUIs should be designed to recognise and respond in a greater variety of languages. That said, many of the points we make could be reinterpreted within this important context of language. Finally, this paper focuses on the speech that is *output* by a VUI. The important argument for VUIs to be designed to recognise a greater variety of speech input, e.g. different accents, is beyond the scope of this paper. To summarise, this paper is about *how VUIs sound*, not what they say, and not what they hear.

[1] In linguistics, the terms speech and voice are used to refer to different things, but this distinction is not important for present purposes and thus the two are used interchangeably.

The paper is structured as follows. First, we introduce the field of sociophonetics and review its main research interests and practices. We then provide a brief overview of the development of devices with voice output to date in order to set the context within which voice output is currently investigated. Through critiquing recent work, we will highlight some of the main areas in which sociophonetics finds points of connection with the design of VUI voices. Next, we look at the ways in which sociophonetics might be integrated into the process of designing VUI voices (whether this is through engaging with sociophonetic literature or having a sociophonetician as a member of the design team is likely to be dependent on the complexity of sociophonetic factors involved in that particular design context and which of the design strategies that we describe herein is being implemented). We primarily illustrate this through discussing three design strategies that are grounded to different degrees on sociophonetic research. These three strategies acknowledge the increasing benefits, and difficulties, encountered as sociophonetics becomes more integrated with voice design. First, we introduce the design strategies *Individualisation* and *Context Awareness*. We offer these two as soft starting points for considering the role that sociophonetics might play in voice design. However, we argue that in order to integrate sociophonetics and voice design to a greater degree, one must consider technolingualism [72] and its relationship with the social construction of technology [8]. We introduce these theories and explain how technological developments will influence speech ideologies (the attitudes and ideas about different speech types) and the potential impact this can have on speakers. Thus, we propose a third design strategy - *Diversification* - which brings into focus the potential consequences of choosing certain voices over others. We close the paper by outlining ways forward for the future at the intersection of sociophonetics and HCI.

3 A PRIMER ON SOCIOPHONETICS

Every voice is unique as a result of the combination of a speaker's physical and social characteristics. Tiny differences in the shape, size, and relative positions of the speech articulators (e.g. teeth, lips, tongue) define how the voice sounds, while the speaker's social history (e.g. where they grew up, the social groups that they belong to) influence how the voice is used [27]. At the same time, sociophonetic work has shown how speech features pattern systematically across social categories. It is these patterns that we introduce in the next section.

To bring our discussion to life, we draw upon well-known personalities from broadcast media. The majority of descriptions below are not technical but attempt to give the impression of speech through spellings. Also, most of the examples are from studies of speakers of English, although there are

bodies of sociophonetic work on other languages. It is most important to bear in mind that this is far from a critical review of sociophonetics; our aim is to introduce the field's work through a selection of accessible examples.

Accent and Voice Quality

Sociophonetics studies two main aspects of speech and voice - *accent* and *voice quality*. An accent is the sum of a set of speech features that a group of people share; it indicates where a person is from, geographically or socially [19]. Voice quality is how a person habitually uses the voice and speech articulators, largely as a result of biology [36], e.g. Donald Trump's[2] voice quality is a result of a fronted tongue, lip rounding, and denasality. How the vocal cords are habitually used is also considered, e.g. Kim Kardashian[3] often has a creaky voice as a result of her vocal chords vibrating slowly.

How we react to voices

Studies have shown how people judge others based on their voice. Within moments of someone speaking a listener may infer where they are from, make other assumptions about their social identity, and apply social stereotypes. For example, a British survey found that the Birmingham accent was the least attractive out of a set of 34 UK English accents [15]. One's social history also influences the social perception of voice. Tompkinson [86] had UK participants listen to sentences, some of them bomb threats, in various accents. The bomb threats were perceived to be more threatening by older participants when spoken with a Northern Irish accent, yet the same effect was not seen in younger participants. This generational difference could be explained by the older participants connecting the Northern Irish accent with the Irish Republican Army, a group that carried out bomb attacks in the UK, mostly in the 1970s and 80s, before the younger participants were born.

Social Categories and voice

Speakers unconsciously use fine phonetic details (i.e. subtle features only identifiable with expert training) to portray and observe aspects of social identity [26, 27]. Our voices are very flexible resources that we can use to signal different social factors at different times, and one aim of sociophonetics is to investigate these practices. That is, how the same person will speak differently within different contexts and with different people. For example, a seminal sociophonetic study is that of 'Heath' [75], who used a falsetto voice to construct a 'diva' persona when with friends, but not with family or at work.

Another aim of sociophonetics is to study the indexical speech features of a group and the associated sociocultural assumptions that they carry. Sometimes these indexical features inform stereotypes about the ways groups of people speak [39]. These social groups may be defined through ethnographic observation (e.g. [25]), social network analysis (e.g. [49]), or by social categories such as geography, sex and gender, age, sexuality, and social class.

Geography. When you first meet someone, you might have an idea of where they are from by their voice. This recognition is usually because of a geographical accent and may be on an international or national scale (i.e. which country, or city they are from). Identifying geographical accents requires experience. Only with exposure to accents are we able to associate them with geography because the connection between location and speech is arbitrary and historical, not iconic. This is true for all social categories and ways of speaking, in fact. Also, new geographical accents develop overtime as a result of continued migration (e.g. [7, 92]. Finally, *everyone* has an accent, although those who speak the "standard" accent of their nation (e.g. US Standard, Received Pronunciation in the UK) may not believe this. The notion of a standard accent and its prestige is purely socio-historical [90] and is caused by these accents being those of dominant social classes, used in the news media, and had their pronunciations taught in schools [30, 37].

Sex and Gender. "[T]he biological effects of speaker sex on speech are in many respects obvious and impossible to avoid" [27, p. 711], the primary difference being longer and thicker vocal chords in males resulting in a lower pitched voice. But techniques can be used to overcome these anatomical constraints, so a speaker may portray their preferred gender (see [48]). Indeed, in sociophonetics the effects of biological sex on the voice are separate from the socially-constructed and performed concepts of gender. Not all men/women speak the same, of course, and differences can be attributed to the kind of man/woman the speaker wants to portray in contrast to other men/women. For example, in Glasgow UK young working-class women produce 's' in a way that is more similar to working-class men than young middle-class women, emphasising local solidarity rather than gender [83].

Age. Language differs from one generation to the next and social commentary frequently includes older generations claiming that young people are destroying or damaging language (as described in [46]). Voice is a key part of language change. One recent change in English is *uptalk*; pitch rises at the end of a statement so it may sound to some listeners like a question. This can now be heard from many people under 30 years old in the Anglo-sphere [89]. Another example is dropping "y" before "oo" in Canadian English, so words like "news" are now more likely to be pronounced "nooz" rather than "nyooz" [13].

[2] www.youtube.com/watch?v=SkHa2-c_8Pk
[3] www.youtube.com/watch?v=R8mcBdBL-t0

Sexuality. Sociophonetics attempts to understand the systematic relationships between speech and social categories, thus sexuality has also been examined, primarily the speech of gay men. This interest has probably been fuelled by the supposed stereotype of the "gay lisp" [40]. This stereotype is most likely due to the 's' sound typically being longer [38, 40, 78] and higher pitched [43, 53] compared to straight speakers. In contrast, few studies have examined gay women's speech [53, 73, 88]. It is important to note that the participants in these studies would have been open about their sexuality and self-identify as part of a gay community. Speech is a learned behaviour; we adopt speech features that signal a community's identity through exposure to that community and a desire to signal affiliation.

Social Class. Speech is often a marker of social class. In Labov's seminal study [34] he went to three department stores in New York City, each one perceived as catering for a different social class of customer. He asked hundreds of staff a question to elicit the answer "fou**r**th floo**r**"; the likelihood of the staff member pronouncing the "r" sounds increased as the social class of the shop increased. Similar results have even been found in societies where class is less engrained. In Beijing, professionals working for foreign companies (a prestigious job) pronounced "r" after a vowel more often than professionals working for state-owned companies [95].

Social Qualities and voice

In most sociophonetic work, speech and its associated social qualities are considered in conjunction with social categories (e.g. masculinity and femininity in studies of gender). As was mentioned earlier, people change their speech to respond to different interaction contexts. It is across these contexts that different social qualities may be invoked by a speaker in their speech. For instance, Kiesling [32] found that members of a college fraternity pronounced the "ing" at the end of verbs as "in" (e.g. "sleepin" and "cookin") at certain times to invoke the qualities of masculinity and physical power. A broader example is that in more formal settings speakers often produce speech that is more similar to the "standard" accent of their nation [35].

Another lens through which we can gain insights into associations between voices and social qualities is voice characterisation; examining what sorts of characters in entertainment media are portrayed with what kinds of voices. Lippi-Green's [41] analysis of Disney films found that all the characters who spoke African-American Vernacular English (AAVE) were animals. Also, most of these characters had questionable social qualities, such as the hedonistic Crows[4] in 'Dumbo'. The only AAVE character that portrayed overtly positive social qualities was the matriarchal Big Mama[5] in 'The Fox and the Hound', although this is still reinforcing a racial stereotype. But it should be highlighted that there has been much change in Disney films in the 20 years since.

Summary

This primer on sociophonetics demonstrates the breadth and depth of the field's inquiry. We now provide a brief overview of the development of VUIs to define the context within which voice-based output has been investigated to date. HCI related work will be critiqued from a sociophonetic perspective to highlight the main areas in which the two fields connect in the design of VUI voices.

4 VOICE-BASED INTERACTION

The first computer synthesised speech was produced at Bell Labs in 1962, and it wasn't long before a series of other labs produced their own. The initial market for computer generated speech devices was people with disabilities. This market can be broadly separated into screen and print readers (using text-to-speech processing) for people with visual impairments and voice output communication aids for people with conditions that prevent speech production [79].

It is as a result of the development of synthesised speech (and the parallel development of automatic speaker recognition) that VUIs are possible. Pearl [71] refers to two VUI eras. The first is defined by over the telephone interactive voice response systems becoming mainstream in the early 2000s. The second is one of Intelligent Personal Assistants that feature VUIs, such as the devices mentioned at the outset of this paper. While this paper is motivated by the recent resurgence of interest in VUI, there are many other devices and contexts within which recorded or computer synthesised voice output is used and our discussion will be relevant (e.g. self-service check outs).

Returning to computer synthesised speech, the range of voice outputs used by the commercially popular Intelligent Personal Assistants is currently very limited. However, diversity may increase over time as knowledge from producing text-to-speech voices is migrated to VUI voices as there is a far greater variety of voice in text-to-speech platforms (e.g. MacOS's Sierra VoiceOver Utility offers 25 voices for screen reading in English). The current interest in VUIs emphasises the need to examine how the sound of a computer-generated voice can affect user experience. The impact of different kinds of voices being used has been minimally explored previously, and with the intent to answer very different research questions. Thus, it is evident that the knowledge and expertise of sociophonetics has so far been untapped. We review this literature below.

[4] www.youtube.com/watch?v=_v2exWrsGOc

[5] www.youtube.com/watch?v=Gy2EOoMua2M

Technology and studies of voice

The vast majority of early studies of voice-based interactions were done to investigate the 'computers are social actors' (CASA) theory. This theory posits that HCI is fundamentally social and so the same social rules and expectations that humans would apply to humans will also be applied to computers. First defined and investigated by Nass et al. [67], their approach was to provide a computer with characteristics associated with humans, namely a "human-sounding" voice, and then establish if participants applied the same social rules as is expected in human-human interaction. The results supported the CASA theory: participants responded to different voices as if they were distinct social actors, and to the same voice as if it were the same social actor, regardless of whether the voice was from the same or different computer. In these studies, the potential effect of the voice was minimised by using several voices for each experiment and randomising them across the participants. Thus, this early work did not consider the sociocultural cues that can be communicated through the voice itself. But the finding that different voices indicate different personas and that this voice-persona association is retained across devices provides a foundation onto which a sociophonetic perspective of VUI voices can be built.

In [67] there is one study where voice was the independent variable. Study 4 compared participants' responses to computers with recorded human male and female voices. The participants' responses evidence the application of human gender stereotypes to computer-embodied voices, e.g. the male voice was rated as more dominant, forceful, and assertive than the female voice. Nass and colleagues continued to explore gender, but began to consider the phonetic features of voices. Nass and Min Lee [64] manipulated speech features (average pitch, pitch range, volume, and speech rate) to produce two computer synthesised voices - one to portray an extrovert personality, the other an introvert personality. They tested these in a 2x2 experiment set up to test whether the listeners (some of whom were defined as introvert, some as extrovert) would rate the same book reviews differently based on hearing the different voices. They found a strong similarity attraction, that is the introvert participants rated the introvert computer voice as more attractive, credible, and informative, while the extrovert participants rated the extrovert voice more highly. Expanding on these findings, it was found that the personality conveyed by the voice was the dominant percept even when the personality conveyed in the linguistic content differed, so introvert texts produced by an extrovert voice made listeners perceive the writer as extrovert [65]. Nass and Min Lee's work [65] highlights the space available for collaboration between HCI and sociophonetics in two ways: 1) in evidencing that social cues in voice override the social cues of linguistic content in devices with voice output; and 2) in making the argument that the socio-psychological processing of speech is the most compelling explanation for the underlying processes that lead to people to apply human social heuristics to computers (CASA), particularly with synthesised speech.

Similarity-attraction effect. An extension to the CASA work has been the investigation of the similarity-attraction theory - that individuals will prefer to interact with others who are similar to them [66]. Although [66] adapted this theory from psychological studies of personality, it appears to have expanded in HCI to be an explanation for other apparent similarity-attraction effects, namely with respect to accent. In [20] Swedish and American participants were interviewed by a computer with voice-output that was either English with a Swedish accent or an American accent. The Swedish participants preferred being interviewed by the Swedish accent than the American accent and rated this voice as more socially rich, and the 'interviewer' as more sociable and likeable. The findings were the opposite for the American participants. Extending this initial work, in [21] participants were asked to rate their experience of interacting with a tourist information website. The website was either about Stockholm or New York and its content was relayed in English with either a Swedish or an American accent. The Swedish participants rated the Swedish English voices as more likeable than the American English voices. Also, the Swedish participants rated the information that the Swedish English voices relayed as more valuable and likeable, even when the information was about New York. The results were again the opposite for the American participants. Their conclusion was that the similarity-attraction effect is so strong it over-rode contextual information.

As the availability of computer synthesised speech has increased and its quality improved, experimental studies have shifted to using these voices rather than recordings of humans. Thus, this work can more directly be related to current commercial VUIs. These studies also reflect a migration of voice-based output from desktop screens to mobile devices and embodied social robots. A New Zealand based study [84] found that users of a healthcare robot had more positive feelings towards it and viewed its performance as more satisfactory when it had a New Zealand accent compared to when it had a US accent. Likewise, a group of children in Ireland preferred the robot that spoke in a UK-accented voice rather than a US-accented voice [80]. In another study in Ireland, Cowan [17] found that users of a navigation system rated the system with an Irish accent as more trustworthy than a US accent irrespective of system accuracy.

While the literature to date shows an awareness of variable features of speech and the impact they might have on

HCI, they only scratch the surface of the full range of variability. It is evident that there are a number of the ways in which sociophonetic knowledge can help direct and shape future investigations. First, voice and speech features can vary across very small groups. In most of the cases described above, "similarity" has been conceptualised as sharing the same nationality with the experiments contrasting a native accent with non-native ones (aside from [80] which does not use an Irish accent). This overlooks the great wealth of geographical variation in accents, not to mention other kinds of social variation (e.g. gender, age, class) that is also likely to interact with this geographical diversity. As [25]'s work in high schools demonstrates, indexing group identity and membership by speech can occur at a micro, hyper-local level. A sociophonetician would view the contrasting of native and non-native accents as an oversimplified method for investigating similarity-effect.

Second, it should be borne in mind that a "standard" accent does not mean the "best" or the most acceptable. Based on their descriptions, it is reasonable to assume that the accents chosen for these studies are those that are viewed as the "standard" of the nation, as discussed earlier (see *Geography*). From their results, [20, p. 300] concludes that "[d]esigners should use the accent that is the most widely accepted as the standard accent within the nation" but this recommendation is presumptuous considering that non-standard and regional accents are yet to be used in such experiments. Further, there is evidence that not everyone views the "standard" accent so positively (e.g. [15]).

There is one study that does not provide evidence to support the similarity-attraction theory, and thus touches on the complexities of voice. This study [68], conducted in Singapore, found the participants rated an over the telephone virtual helpdesk assistant as politer when it used English spoken with a British accent rather than with the local Singaporean accent. Interpreting this result requires knowledge of the sociocultural context surrounding speaking English in Singapore. The authors of [68] state that the voice with a Singaporean accent resembled a variety officially referred to as "Singaporean Colloquial English" and colloquially known as "Singlish". This variety is regularly used in casual contexts but is avoided in the workplace and formal contexts. Hence, the use of this accent within a context that it is not typically associated with (the formal context of an information giving service from a company) is less likely to receive a positive response from participants. An alternative explanation was also given by the authors; that many in Singapore view British culture as prestigious as a consequence of Singapore being historically a British Crown Colony. Thus, the British accent's prestige resulted in a more positive response from the participants. Regardless of the explanation, this result evidences the need to account for the sociocultural attitudes and contexts that may be present when designing and evaluating voice output, as the authors conclude.

Summary

To conclude, investigation of how social cues in the voice can affect user experience has been restricted to relatively few studies on gender and native/non-native accents. The challenge here, from a sociophonetic perspective, is that this leads to oversimplifications of the sociocultural complexities of voice, and overlooks the many other features of voice that can vary systematically and meaningfully. However, these findings show that it only takes a few cues to be present for humans to respond to computer generated speech as though it is a social actor. This implies that more detailed and complex layers of sociocultural cueing in computer generated speech may also be responded to as though its source was a human. The importance of this in HCI terms is to suggest that there are much richer landscapes of design possibility when we have a more nuanced understanding of the role of voice in interaction. This lays the foundations for considering the design of VUI voices from a sociophonetic perspective.

5 INTEGRATING SOCIOPHONETICS INTO VOICE DESIGN: INITIAL DESIGN STRATEGIES

In this section, we define and explore two initial design strategies: i) *Individualisation*; and ii) *Context Awareness*. We introduce these as the starting points that designers could take in attending to sociophonetic aspects of voice. *Individualisation* follows current common approaches of designing one voice output per assistant but suggests that users should have greater choice in the voice their assistant produces. *Context Awareness* instead results in multiple voices being heard through the VUI depending on its use and context. As previously stated, while both [14] and [29] consider the voice in VUI design they give little advice beyond this signpost. Also, while the two design goals given here are also identified in [62, 63] their recommendations were motivated by principles of evolutionary psychology, not those of sociophonetics. Plus, we expand upon this previous work significantly in providing a range of ideas on how these design goals might be met considering today's technology.

Designing for Individualisation

The first design strategy is to allow for *voice output to be individualised*, as has occurred with many other IoT and mobile devices. To reiterate our earlier discussion, one's social history may result in associating certain kinds of voices with specific emotions and experiences. This can be within a national context (e.g. [86]), or an individual's life experiences. For example, one may associate a certain voice quality with a boss who was difficult to work with, or an accent with a regular childhood holiday destination. Thus, in this design

strategy a basic understanding of sociophonetics is incorporated; that people have different responses to voices based on personal experience. We can imagine individualisation enriching user experience in a multitude of ways. For example, in behaviour change contexts users may feel more committed to following guidance or feedback when it is produced in a voice the user feels a positive affinity towards. Equally, producing such content in a voice with qualities that the user associates with people that they have previously received directions from (such as a teacher or boss) may actually result in higher rates of response and behaviour change over a shorter period of time.

It is possible to imagine a series of methods by which this individualisation could occur, each with their own nuances. To incorporate voices the user has a positive affinity to, users could select a voice from a suite of options such as is already possible with in-vehicle satellite navigation systems with voice output [87]. Nonetheless, the number of voices in current VUI systems is very limited. Beyond this, it is not hard to imagine a service where a user can select and combine different components of the voice (e.g. accent, voice quality, speed), and then even manipulate these to be more finely tuned to their preferences (e.g. selecting the degree of 'roughness' or 'softness' for the voice from along a continuum). This would allow for the voice of the VUI to be based on each user's preferences, which will probably reflect the user's unique social history. Taking the approach where the voice is dependent on the user's selection alone requires minimal engagement with sociophonetic literature apart from understanding its basic concepts. Alternatively, preferences could be collected by other means, such as listening to voice clips and then rating them based on different qualities, just like the set-up of a typical sociophonetic perception experiment.

An alternative to explicit exploration and selection of a voice could be "voice matching". In a similar vein to dating websites, a user could answer a series of questions and an algorithm could recommend several voices from a larger selection. Further, the individualisation of the voice could be conducted by the assistant system/platform as it tracks how it is used. Presumptions about the user's age, class, gender, and personality type can be inferred from a multitude of activity data, such as the music listened to, the movies and TV watched, the restaurants and shops visited, and the items purchased from them. What could satisfy the similarity-attraction effect more than a voice output based on one's own personal data? This method would be most appropriate for VUI assistants that are available across mobile and stable devices, so that a greater range of activities can be pooled into one repository to give a more detailed representation of the user's lifestyle. Where a user's preferences in regards to voice is inferred through social categories via different data types greater engagement with sociophonetic literature is required to be able to identify the speech features that index such information within that user's cultural context and ensure these are present in the voice that is selected.

Most of these methods would be relatively straightforward to implement because they are inspired by those found in other contexts (e.g. satnavs, dating websites). However, integrating sociophonetics with voice design in this relatively simplistic manner will reap few, limited benefits, and actually raises more complex issues that we discuss later on.

Designing for Context Awareness

This second design strategy considers *voice design in relation to the contexts within which it is used.* An obvious example would be geographical accent. Many intelligent personal assistants are location aware (to allow for weather and travel information, among other functions), thus an initial step could be the VUI assistant producing an accent that reflects the geographical location. This could be based on broader, large areas or specific, smaller locations. Consider the UK. The default position seems to be to provide UK VUIs with a "UK" voice that speaks with a Southern English accent. Such a voice is not representative of Wales, Scotland, or Northern Ireland, and a VUI producing a voice to represent these large, but distinct areas of the country could be an initial step to reflecting geographical location. Finer grained personalisation would be for the VUI voice to represent the city it is located in, for example the accents of Glasgow and Edinburgh are distinct although both are Scottish.

In order to take this design sensitivity further, it would be essential to allow for multiple voices in one VUI device and for these voices to be tailored to the context, predominantly the activities that are being performed. As we know from Nass et al. [67], people respond to different voices as if they were different social actors, even if they are output from the same device. Current VUI assistants can be used to perform a range of tasks, yet the approach thus far has been for one device to exhibit one voice. For the technology companies producing these VUI assistants this is logical not just from a technology development point of view but also from a branding perspective; the voice represents the technology company and their product, just like a company's logo. Thus, we are not proposing these default voices be removed, just for additional ones to be added.

As is evident from sociophonetic literature, voices differ across a range of social categories and social qualities, and the social aspects are communicated to the listener through the voice's features. This social cueing could be harnessed in the design of VUI voices in a number of ways. For example, when interacting with a VUI to conduct banking activities (e.g. pay a credit card bill) engendering the VUI's voice to portray certain social qualities may enhance the experience in this interaction. If we were performing such tasks with

a real person, what social qualities would we want them to embody? Probably trustworthiness, honesty, and efficiency. Thus, ensuring the voice used in the interaction is perceived to embody these social characteristics is likely to enhance the user's perception of a successful interaction.

The voice could also complement activities within the context of entertainment and leisure time. For example, a user may have a particular genre of music they habitually listen to. This is already monitored by music streaming services (e.g. Spotify) and so could be acknowledged during interaction with a VUI. What if a user's favourite genre of music was Reggae? Or Country? For each of these genres we can imagine a very different voice, or even a famous artist's voice that is particularly associated with that genre. This is because music genres often originate from a specific location at a specific time in history, hence each can be associated with a particular accent that is flavoured with the sociocultural context of that time. Extending this scenario further to IoT devices, another example could be monitoring the choice of television programs and movies played through a TV that is integrated with the home's VUI assistant. Imagine what a VUI's voice might sound like if it were tailored to introduce a horror, western, gangster, or British romantic-comedy films.

Thus, in this design strategy an understanding of how people have different responses to voices based on personal experience is necessary, but so is an understanding of what aspects of voice trigger these responses and why. Further, the design of these voices needs to be explicitly placed within a context that is defined by geographical and social factors, and not just an individual's personal history, and negotiating each of these elements in voice design may be difficult.

Summary

These two design strategies indicate the wealth of opportunities that sociophonetics brings to voice design. But these are still very simple starting points, and as such are relatively crude. While drawing on individualisation and context awareness would extend the current range of possibilities in voice design, the sociocultural aspects of voice would still be relatively trivialised. In the following section we will explain the root of these concerns by introducing technolingualism [72], a theory that, we argue, supports a more nuanced integration of sociophonetics with voice design. After discussing technolingualism, we present the third design strategy - *diversification.*

6 REFLECTING ON TECHNOLINGUALISM

Pfrehm defines technolingualism as the phenomenon that "technology both shapes and is shaped by language" [72, p. x], the term 'language' being used broadly to refer to communication resources (speech, verbal language, written language, etc.). Let's take the telephone as an example. The telephone was shaped by language because its design was based on the physical properties of speech. The telephone shaped language by influencing practices and ideologies. The need for language to indicate the start and the end of a conversation was not required in face-to-face interactions but became necessary on the telephone. Hence, greetings (e.g. 'hello') and farewells (e.g. 'goodbye') had to develop. Thus, ideologies evolved of what is and is not appropriate telephone etiquette. Over time, greetings and farewells began to be used in face-to-face interactions, and this shaped interaction etiquette in this non-technologically mediated context.

The concept of technolingualism shares some qualities with the wider theories and concerns of the social construction of technology (e.g. [8]). Bringing these ideas together, we start to see the great complexity of how technology, voice, society, and design are enmeshed and inform one-another. To elaborate: people shape language which in turn influence peoples' language related experiences; and people bring these language experiences to the design of technology, and that technology then shapes subsequent language experiences. To more clearly illustrate how technolingualism in the context of voice output design would occur, we describe it as four inter-related processes: i) *People shape language*; ii) *Speech ideologies shape people's experiences*; iii) *People shape technology*; and iv) *Technology shapes language.*

People shape language

As was explained earlier, only with experience are we able to associate speech features with social categories (location, gender, age, etc.) because the connection between these factors and speech is arbitrary and historical, not iconic. Thus, through encountering speakers and their way of speaking people associate certain types of speech with social categories and social qualities. From this, individual (that is personal) and collective (that is at a local, national, or international scale) speech ideologies develop. Speech ideologies are the attitudes and ideas about different speech types. Take the idea of the "standard accent" and how a particular way of speaking becomes the standard: A group with some sort of power (social or economic but most likely both) possess a shared way of speaking that becomes indexical of that group and associated with their power. It thus takes on positive evaluation across a population. In parallel, negative connotations become assigned to other ways of speaking. These ideologies may become further established via contexts such as education and employment, especially since the socially powerful group has disproportionate influence over how language is used and evaluated in such context. So, speech ideologies are created and enforced by people [50].

Speech ideologies shape people's experiences

Speech ideologies can in turn lead to accent prejudice and accent-based discrimination (as investigated in [15]). This is when a preconceived, often negative, opinion about someone based on how they speak (as a result of that speech's connection with the social background of the speaker) is acted upon [70, p. 127]. For example, the British MP Angela Rayner regularly receives online abuse because of her Manchester UK accent, e.g. saying she sounds "thick" (unintelligent) [24, 31]. Accent prejudice is rarely talked about in comparison to the prejudices held against genders, sexualities, ethnicities, or social classes; although accent prejudice is almost always as a result of holding these other prejudices. This may be because people are unaware that the biases or prejudices that they hold can be activated by hearing voices, or the lack of condemnation of accent prejudice could result in the perception that it is acceptable. Thus, unlike other kinds of discrimination, we have no statistics for incidences of discrimination because of speech. But there are many case studies and personal accounts (for examples see [1]), and sociophonetic studies continually reveal the likely consequences of discrimination based on the way someone speaks: [23] found speakers with non-standard accents were perceived as guiltier, which could have consequences within court proceedings; [77] found non-standard speakers are less credible in radio advertisements; and [6] found UK teachers with regional accents are being told to sound more 'professional'. The study of non-native accents is more complex although such speakers were judged as less employable, particularly for customer facing roles, in [85].

People shape technology

As is outlined by the social construction of technology [8], technology does not determine human action but the design decisions that humans make do shape behaviours that occur as a result of using technology. With voice output, and in reference to the example of the standard accent, designers and produces of technology are defining what is and is not accepted as 'good' or 'appropriate' speech. So far, the "standard" speech for VUIs is also the "standard" for their national context; the accent and voice quality that is encouraged within the education system, has historically been dominant in the media, and used by the upper class. Thus, the prototypical voices of VUIs are already engrained with the speech biases of their designers. Some may argue that this is not as a result of bias. There may be more knowledge of how to generate these sorts of voices, or the designers may have considered that these voices would be the most intelligible because they are encouraged in the education system. But these responses in fact further evidence socio-historical biases for and against certain speech types.

Technology shapes language

As should now be evident, the current use of certain types of voices over others in VUIs is translating already engrained speech ideologies into this new interaction context. This reinforces speech ideologies (i.e. that a certain way of speaking is the best or most prestigious, or that others are incorrect or not appropriate). From speech ideologies come prejudices against ways of speaking, which can result in discrimination against those speakers.

Summary

It should now be evident that VUI interactions do not occur in a vacuum but are informed by prior interactions. Thus, VUI interactions will inform future interactions, both with VUIs and in other contexts also. Therefore, voice output does not just influence the interaction context that it has been designed for, but subsequent interactions in other contexts as well. But unlike other technological revolutions that just placed language in a new medium and interaction context (e.g. writing, the printing press and type writer, the telegram and telephone) the voice output of a VUI can be viewed as a new communication partner, as Nass's [64, 65, 67] work evidences. This new partner is an amalgamation of the ideologies of many stakeholders (e.g. Natural Language Processors, Speech synthesisers, Ix designers), and so is also an amalgamation of many potential communication partners. Hence, we argue that sociophonetic integration into voice design will be limited unless engrained speech ideologies are highlighted and considered in the design process.

7 REFLECTING ON OUR INITIAL DESIGN STRATEGIES

Now technolingualism has been introduced, it is possible to more fully communicate how and why our initial design strategies raise complex issues. In the *individualisation* design strategy, we make two suggestions: i) that a user can select the voice they would like for their device through a range of means, and ii) that the voice output could be personalised to reflect the user's social categories by interpreting a multitude of usage data. Both of these overlook the sociocultural complexities of the voice. First, it is reasonable to predict that a user's selection of a voice will reflect prejudices both at a personal and at a societal level, e.g. users may select voices that reflect their nation's "standard" way of speaking, reiterating the historical attribution of prestige. Further, producing a voice based on one's own demographics will create an echo. We know that for groups of people to sound like each other they must spend time interacting. Many people's experiences of voice diversity in face-to-face interaction is limited to those that sound like them. Thus, it is reasonable to assume that exposure to different types of

speech is predominantly through mediated communication (e.g. radio, television). Prejudice in general partly comes from a lack of exposure to difference, and the four dominant theories that relate knowledge, stereotyping, and prejudice all posit that healthy relations require a high degree of in-group and out-group communication [44]. We do not argue that the design of voice output will resolve prejudices, but this context does provide an opportunity to increase exposure of speech diversity, which may contribute to its acceptance.

The second design strategy - *context awareness* - also overlooks the complexities of voice and its potential consequences somewhat. Again, we make two suggestions i) to design voices based on social categories in relation to activity, and ii) to design voice in relation to social qualities. Designing a voice based on social categories, without the expertise of a sociophonetician who specialises in that particular type of speech, there is the danger that exaggerated, caricature-like voices will be produced. These exaggerations will probably be disrespectful to genuine speakers and could even encourage prejudices and negative ideologies by making them appear acceptable. This point is relevant for any voice design, regardless of whether the voice is designed for a particular activity as we suggest in this design goal. Similarly, designing voices based on social qualities could also reinforce current prejudices. As was outlined earlier, the perceived social qualities of a group become attached to the way that they speak. It is not that social qualities become associated with ways of speaking directly. Therefore, designing a voice to portray certain social qualities will tap into, and subsequently reinforce, the stereotypes associated with a social group.

Designing for Diversification

Following our discussion of technolingualism and the sociocultural shaping of voice technology, and our reflections on our initial design strategies, we introduce a further design strategy - that of *diversification*. We introduce this to encourage the critical consideration of the sociocultural aspects of voice in detail (as [81] touched upon) with particular focus on the potential consequences of design decisions, and advocates for the active engagement of sociophoneticians in decision making.

To put it simply, diversification primarily calls for voice output to deviate from the perceived "standard" of nations. Incorporating different types of voices that are representatives of a greater variety of social groups begins to both suppress the prestige that the national standard receives and alleviates some of the prejudices that non-standard ways of speaking are put under. Of course, this would have to be approached with sensitivity to avoid voices that still embody negative stereotypes and prejudices (see the example of Big Mama in [41]). Avoiding the reinforcement of speech ideologies, and challenging existing ideologies are two different things, however. The question should thus be raised about which of these approaches to take, and how.

Preventing the reinforcement of speech ideologies. In order to prevent the reinforcement of current speech ideologies, we imagine four strategies, all of which require the expert involvement of a sociophonetician. First, while the majority of this paper's primer on sociophonetics is about the speech and voice features that index social groups, not all features are associated with social categories or qualities. For example, voice quality is predominantly related to biological differences [36]. As a starting point, this can be utilised to enable the design of a selection of voices that are different from each other, but the perception of difference is not activated by using sociocultural knowledge in the listener.

Second, not all accents will trigger associations with social categories and social qualities, or stereotypes and prejudices, in every listener. As we have explained, the connection between an accent and the social information it is indexes (location, gender, sexuality etc) develops through experience. If a listener is unfamiliar with a certain accent or way of speaking, these will not be associated with sociocultural information to be activated when listening to the voice. Hence, one could take the *individualisation* design goal that we have proposed and reverse it; understanding a user's personal history to identify voices and ways of speaking that are unfamiliar, but still intelligible, to them.

A third strategy, that closely aligns with the strategy above, is to invent new accents. To review our discussion earlier, an accent is the sum of a set of speech features [19] and there are hundreds of ways that the voice and speech can differ and be manipulated. Thus, sociophoneticians could invent new accents to be used in VUIs, just like new ways of speaking are invented as a part of conlangs (languages that were invented rather than naturally evolved [69] such as Esperanto, Elvish, Parseltongue, and Klingon). Such an accent would be designed to have no (or few) prior associations with social categories or qualities and would therefore be less tainted with stereotypes and prejudices.

Finally, reducing the human-likeness of VUI voices could circumvent the triggering of social stereotypes. Some argue that human voices are inappropriate for VUIs from a usability perspective (e.g. [51, 52]). Such arguments are motivated by research questions around the alignment of user expectations with technological possibilities of not just VUIs but conversational interfaces in general (see [9, 16, 42]). However, such a strategy could also address sociophonetic concerns in regards to voice, and that an initial finding is that the reduction of human-likeness of voice does not necessarily lead to differing perceptions of qualities such as appeal and

credibility [11] suggests this approach should be investigated further.

Challenging negative stereotypes. The design sensitivity of *diversification* requires more nuance. If an adversarial design approach was used [22] then the selection of a device's voice would come as a result of finding out the groups of people (e.g. nationalities, ethnicities, sexualities) that the user dislikes or disagrees with, and then provide voices that represent these groups. This would lead to technological interactions that the user will find inappropriate, unusual, or unpleasant. Rather, we suggest that designers of VUIs consider a range of critical questions to ask themselves to promote sociophonetic sensitivity and more reflective voice design practices: Could this voice privilege certain people over others?, Could this voice be viewed as racist/sexist?, How might someone who speaks in this way react to this voice being used in this context?. This could partly mitigate for the lack of a sociophonetician in the design process, although considering the complexity of sociophonetic concepts being considered within the *designing for diversification* strategy it is highly likely that the active engagement of a sociophonetician would be required rather than merely consulting sociophonetic literature.

To conclude, the design and selection of voices for VUIs is complicated and sensitive. The ramifications of design decisions could go far beyond the immediate impact of interaction with the device. It is evident that in the design of VUI voices there is a social responsibility to be respectful of the vocal diversity found across the world and to avoid contributing to the othering of social groups.

8 WAYS FORWARD

Talking about speech and voice is actually very difficult. Without specific training in phonetics, the finer details of the speech signal are processed unconsciously, and yet the sociocultural information in the voice is inferred without the listener's awareness. Hence, unless a listener's attention is specifically drawn to a certain voice feature, its presence and role as a conduit for sociocultural information is unappreciated. Indeed, in writing this paper we have had to rely on the voices of media personalities in an attempt to provide accessible illustrations. Also, when accent features or voice quality are noticed, they are difficult to describe unless the hearer is party to technical terminology. Finally, just like with other sound mediums, voice is fleeting and intangible. Of course, these issues are not just relevant to users or participants but also to others involved in the design process. These issues are three key areas of future work; i) how can attention be drawn to features of voice; ii) how can voice be conceptualised and discussed in an accessible manner; and iii) how can the social and cultural nuances of voices be acknowledged and interpreted in VUI voice design. Addressing these issues would help to establish voice as a material with which we can design.

To establish an initial avenue that could be taken in the journey to integrate the fields of HCI and sociophonetics, we have considered the design approaches already used in the HCI field. The associations between voices and social categories is historical, and so our perceptions of speakers are based on our cultural experiences of voice. Therefore, sociophonetic knowledge aligns with experience-centred design (ECD); an approach that aims to address people's desires, values and feelings through enriching interaction with technology by considering the personal narratives that they bring to an interaction [93]. In line with the view of ECD, the interests and findings of sociophonetics highlights how one cannot 'design an experience'. Individuals bring their experiential histories (that is their sociocultural knowledge of voice) to interactions, thus one can only 'design *for* an experience'.

This paper's orientation to voice output, and not the processing of speech input, is thus also an orientation to aesthetics. McCarthy and Wright [45] advocate for focus on the aesthetics of an interaction rather than just the ergonomics. The intention here is to make interactions more socially meaningful. Also, this framework shows strategies to individual, human experience. One of the four threads of experience in the framework is particularly attuned to voice from a sociophonetic perspective. The *Sensual* thread draws attention to people's visceral responses to an experience and since we as listeners unconsciously observe aspects of social identity through fine phonetic detail our responses to voices can be considered visceral.

In summary, the research space that emerges from our propositions also needs to focus on developing tools and techniques that allow the mining of and inference of voice-based preferences, while also dynamically respond to users.

9 CONCLUSION

VUIs have achieved such a level of technical capability that attention can move towards considering aesthetics in VUI design. And so, now would be an appropriate time to explore relevant knowledge from other disciplines. Thus, we make an argument for the importance of sociophonetics and how as a research field its expertise may be brought into HCI to support the design of VUI voices. However, as we have explored in this paper, transferring knowledge from sociophonetics into VUI design will be complicated, raises various challenges that need to be addressed, and points to some interesting research questions. We conclude by proposing that voice output is designed to encourage diversity, that we deviate from the supposed national standard at a minimum, and suggest ways in which the bridge between sociophonetics and HCI could be built.

Acknowledgements

We thank Mark Blythe and Abigail Durrant for their comments on a draft of this paper. We also thank our CHI 2019 reviewers for their recommemdations and also our CHI 2018 reviewers who saw a much earlier version of this paper - your critique progressed our work significantly.

REFERENCES

[1] Accentism Project. 2018. Accentism Project. https://accentism.org/

[2] Adapt Centre. 2018. ADAPT Associate Professor Sole Irish Recipient of the Google Faculty Research Award. https://www.adaptcentre.ie/news/adapt-associate-professor-sole-irish-recipient-of-the-google-faculty-resear

[3] Amazon. 2018. Echo & Alexa Devices. https://www.amazon.co.uk/b?ie=UTF8&node=14100223031

[4] Apple. 2018. Siri. https://www.apple.com/siri/

[5] Matthew P. Aylett, Per Ola Kristensson, Steve Whittaker, and Yolanda Vazquez-Alvarez. 2014. None of a CHInd. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI EA '14*. ACM Press, New York, New York, USA, 749–760. https://doi.org/10.1145/2559206.2578868

[6] Alex Baratta. 2017. Accent and Linguistic Prejudice within British Teacher Training. *Journal of Language, Identity & Education* 16, 6 (nov 2017), 416–423. https://doi.org/10.1080/15348458.2017.1359608

[7] Allan Bell and Andy Gibson. 2008. Stopping and Fronting in New Zealand Pasifika English. *University of Pennsylvania Working Papers in Linguistics* 14, 2 (2008), 44–53. https://repository.upenn.edu/pwpl/vol14/iss2/7/

[8] Wiebe E. Bijker, Thomas P. Hughes, and Trevor J. Pinch. 1987. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. MIT Press, Cambridge, Massachusetts.

[9] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, Janet F. McLean, and Ash Brown. 2011. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition* 121, 1 (oct 2011), 41–57. https://doi.org/10.1016/J.COGNITION.2011.05.011

[10] Hadumod Bussmann. 1998. 'speech'. In *Routledge Dictionary of Language and Linguistics*. Taylor & Francis, London, UK, 1106.

[11] Paulo João Cabral, Benjamin R. Cowan, Katja Zibrek, and Rachel McDonnell. 2017. The Influence of Synthetic Voice on the Evaluation of a Virtual Character. In *Proceedings of INTERSPEECH 2017, the 18th conference of the International Speech Communication Association*. 229–233. https://doi.org/10.21437/Interspeech.2017-325

[12] Leigh Clark, Phillip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, and Benjamin Cowan. In press. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* (In press). http://arxiv.org/abs/1810.06828

[13] Sandra Clarke. 2006. Nooz or nyooz?: The Complex Construction of Canadian Identity. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 51, 2-3 (nov 2006), 225–246. https://doi.org/10.1017/S0008413100004084

[14] Michael H. Cohen, James P. Giangola, and Jennifer Balogh. 2004. *Voice user interface design*. Addison-Wesley Publishers.

[15] Nikolas Coupland and Hywel Bishop. 2007. Ideologised values for British accents. *Journal of Sociolinguistics* 11, 1 (feb 2007), 74–93. https://doi.org/10.1111/j.1467-9841.2007.00311.x

[16] Benjamin R. Cowan, Holly Branigan, Habiba Begum, Lucy McKenna, and Eva Szekely. 2017. They Know as Much as We Do: Knowledge Estimation and Partner Modelling of Artificial Partners. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society (CogSci2017), London, UK, 26-29 July 2017*. 1836 – 1841. https://mindmodeling.org/cogsci2017/papers/0355/index.html

[17] Benjamin R. Cowan, Derek Gannon, Jenny Walsh, Justin Kinneen, Eanna O'Keefe, and Linxin Xie. 2016. Towards Understanding How Speech Output Affects Navigation System Credibility. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16*. ACM Press, New York, New York, USA, 2805–2812. https://doi.org/10.1145/2851581.2892469

[18] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What can i help you with?". In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '17*. ACM Press, New York, New York, USA, 1–12. https://doi.org/10.1145/3098279.3098539

[19] David Crystal. 2008. *A dictionary of linguistics and phonetics* (6th ed.). Blackwell Publishing, Oxford, UK.

[20] Nils Dahlbäck, Seema Swamy, Clifford Nass, Fredrik Arvidsson, and Jörgen Skågeby. 2001. Spoken Interaction with Computers in a Native or Non-Native Language - Same of Different?. In *Proceedings of INTERACT 2001 - International Conference on Human-Computer Interaction, Tokyo, Japan, July 9-13, 2001*. 294–301.

[21] Nils Dahlbäck, QianYing Wang, Clifford Nass, and Jenny Alwin. 2007. Similarity is more important than expertise. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*. ACM Press, New York, New York, USA, 1553. https://doi.org/10.1145/1240624.1240859

[22] Carl DiSalvo. 2012. *Adversarial design*. MIT Press, Cambridge, Massachusetts.

[23] John A. Dixon, Berenice Mahoney, and Roger Cocks. 2002. Accents of Guilt? *Journal of Language and Social Psychology* 21, 2 (jun 2002), 162–168. https://doi.org/10.1177/02627X02021002004

[24] Rob Drummond. 2016. Leave off, will you? Britain should celebrate 'regional' accents. http://theconversation.com/leave-off-will-you-britain-should-celebrate-regional-accents-67952

[25] Penelope. Eckert. 1989. *Jocks and burnouts : social categories and identity in the high school*. Teachers College Press, New York, USA.

[26] Penelope Eckert. 2008. Variation and the indexical field1. *Journal of Sociolinguistics* 12, 4 (sep 2008), 453–476. https://doi.org/10.1111/j.1467-9841.2008.00374.x

[27] Paul Foulkes, James M. Scobbie, and Dominic J. L. Watt. 2010. Sociophonetics. In *Handbook of Phonetic Sciences* (2nd ed.), William J. Hardcastle, John Laver, and Fiona E. Gibbon (Eds.). Blackwell, Oxford, UK, 703–754.

[28] Google. 2018. Google Assistant. https://assistant.google.com/{#}?modal{_}active=none

[29] Randy Allen. Harris. 2005. *Voice interaction design : crafting the new conversational speech systems*. Morgan Kaufmann Publishers, San Fransisco, California, USA.

[30] Arthur Hughes, Peter Trudgill, and Dominic J. L. Watt. 2013. *English accents & dialects : an introduction to social and regional varieties of English in the British Isles* (5th ed.).

[31] Oliver Kamm. 2017. Lay off the yod-droppers – all accents are 'bootiful'. https://www.thetimes.co.uk/article/oliver-kamm-the-pedant-695wj30hk

[32] Scott Fabius Kiesling. 1998. Men's Identities and Sociolinguistic Variation: The Case of Fraternity Men. *Journal of Sociolinguistics* 2, 1 (feb 1998), 69–99. https://doi.org/10.1111/1467-9481.00031

[33] Simon King. 2015. A Reading list of recent advances in speech synthesis. In *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK, Paper number 1043. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS1043.pdf

[34] William Labov. 1997. The Social Stratification of (r) in New York City Department Stores. In *Sociolinguistics: A Reader*, Nikolas Coupland and Adam Jaworski (Eds.). Springer, New York, USA, 168–178.

[35] William Labov. 2001. The anatomy of style-shifting. In *Style and Sociolinguistic Variation*, Penelope Eckert and John Rickford (Eds.). Cambridge University Press, Cambridge, UK, 85–108.

[36] John. Laver. 1994. *Principles of phonetics.* Cambridge University Press, Cambridge, UK.

[37] Dick Leith. 1997. *A social history of English* (2nd ed.). Routledge, Oxon, UK.

[38] Erez Levon. 2007. Sexuality in context: Variation and the sociolinguistic perception of identity. *Language in Society* 36, 4 (2007), 533–554. https://doi.org/10.1017/S0047404507070431

[39] Erez Levon. 2014. Categories, stereotypes, and the linguistic perception of sexuality. *Language in Society* 43, 5 (2014), 539–566. https://doi.org/10.1017/S0047404514000554

[40] Sue Ellen Linville. 1998. Acoustic correlates of perceived versus actual sexual orientation in men's speech. *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP)* 50, 1 (1998), 35–48. https://doi.org/10.1159/000021447

[41] Rosina Lippi-Green. 1997. Teaching children how to discriminate: What we learn from the Big Bad Wolf. In *English with an Accent: Language, ideology, and discrimination in the United States.* Routledge, New York, USA, 79–103.

[42] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA". In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16.* ACM Press, New York, New York, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[43] Sara Mack and Benjamin Munson. 2012. The influence of /s/ quality on ratings of men's sexual orientation: Explicit and implicit measures of the 'gay lisp' stereotype. *Journal of Phonetics* 40, 1 (jan 2012), 198–212. https://doi.org/10.1016/J.WOCN.2011.10.002

[44] Jonathan Matusitz. 2012. Relationship between knowledge, stereotyping, and prejudice in interethnic communication. *Journal of Tourism and Cultural Heritage* 10, 1 (2012), 89–98. https://doi.org/10.25145/j.pasos.2012.10.008

[45] John McCarthy and Peter Wright. 2004. *Technology as experience.* MIT Press, Cambridge, Massachusetts.

[46] John McWhorter. 2013. Txtngiskillinglanguage. JK!!! https://www.ted.com/talks/john_mcwhorter_txtng_is_killing_language_jk

[47] Microsoft. 2017. Cortana. https://www.microsoft.com/en-gb/windows/cortana

[48] Matthew Mills and Gillie Stoneham. 2017. *The voice book for trans and non-binary people : a practical guide to creating and sustaining authentic voice and communication.* Jessica Kingsley Publishers, London, UK.

[49] James Milroy and Lesley Milroy. 1978. Belfast: change and variation in an urban vernacular. In *Sociolinguistic patterns in British English*, Peter Trudgill (Ed.). University Park Press, USA, 19–36.

[50] James Milroy and Lesley Milroy. 2012. *Authority in Language: Investigating Standard English.* Taylor & Francis, Oxon, UK.

[51] Roger K. Moore. 2017. Appropriate Voices for Artefacts: Some Key Insights. In *Proceedings of the 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR) Skvöde, Sweden, 25-26 Aug 2017.* 7–11. http://vihar-2017.vihar.org/assets/vihar2017_proceedings.pdf

[52] Roger K. Moore. 2017. Is Spoken Language All-or-Nothing? Implications for Future Speech-Based Human-Machine Interaction. In *Dialogues with Social Robots: Enablements, Analyses. and Evaluation*, Kristina Jokinen and Graham Wilcock (Eds.). Springer, Singapore, 281–291.

[53] Benjamin Munson, Elizabeth C. McDonald, Nancy L. DeBoe, and Aubrey R. White. 2006. The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics* 34, 2 (apr 2006), 202–240. https://doi.org/10.1016/J.WOCN.2005.05.003

[54] Cosmin Munteanu, Ben Cowan, Keisuke Nakamura, Pourang Irani, Sharon Oviatt, Matthew Aylett, Gerald Penn, Shimei Pan, Nikhil Sharma, Frank Rudzicz, and Randy Gomez. 2017. Designing Speech, Acoustic and Multimodal Interactions. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17.* ACM Press, New York, New York, USA, 601–608. https://doi.org/10.1145/3027063.3027086

[55] Cosmin Munteanu, Matt Jones, Sharon Oviatt, Stephen Brewster, Gerald Penn, Steve Whittaker, Nitendra Rajput, and Amit Nanavati. 2013. We need to talk. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13.* ACM Press, New York, New York, USA, 2459. https://doi.org/10.1145/2468356.2468803

[56] Cosmin Munteanu, Keisuke Nakamura, Kazuhiro Nakadai, Pourang Irani, Sharon Oviatt, Matthew Aylett, Gerald Penn, Shimei Pan, Nikhil Sharma, Frank Rudzicz, and Randy Gomez. 2016. Designing Speech and Multimodal Interactions for Mobile, Wearable, and Pervasive Applications. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16.* ACM Press, New York, New York, USA, 3612–3619. https://doi.org/10.1145/2851581.2856506

[57] Cosmin Munteanu and Gerald Penn. 2015. Speech-based Interaction. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '15.* ACM Press, New York, New York, USA, 2483–2484. https://doi.org/10.1145/2702613.2706679

[58] Cosmin Munteanu and Gerald Penn. 2016. Speech-based Interaction. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16.* ACM Press, New York, New York, USA, 992–995. https://doi.org/10.1145/2851581.2856689

[59] Cosmin Munteanu and Gerald Penn. 2017. Speech-based Interaction. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17.* ACM Press, New York, New York, USA, 1196–1199. https://doi.org/10.1145/3027063.3027117

[60] Cosmin Munteanu, Gerald Penn, Cosmin Munteanu, and Gerald Penn. 2014. Speech-based interaction. In *Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems - CHI EA '14.* ACM Press, New York, New York, USA, 1035–1036. https://doi.org/10.1145/2559206.2567826

[61] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18.* ACM Press, New York, New York, USA, 1–7. https://doi.org/10.1145/3173574.3173580

[62] Clifford Nass and Scott Brave. 2005. *Wired for speech : how voice activates and advances the human-computer relationship.* MIT Press, Cambridge, Massachusetts.

[63] Clifford Nass and Li Gong. 2000. Speech interfaces from an evolutionary perspective. *Commun. ACM* 43, 9 (sep 2000), 36–43. https://doi.org/10.1145/348941.348976

[64] Clifford Nass and Kwan Min Lee. 2000. Does computer-generated speech manifest personality? an experimental test of similarity-attraction. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '00.* ACM Press, New York, New York, USA, 329–336. https://doi.org/10.1145/332040.332452

[65] Cifford Nass and Kwan Min Lee. 2001. Does Computer-Synthesized Speech Manifest Personality? Experimental Tests of Recognition, Similarity- Attraction, and Consistency Attraction. *Journal of Experimental Psychology: Applied* 7, 3 (2001), 171–181. http://psycnet.apa.org/doiLanding?doi=10.1037%2F1076-898X.7.3.171

[66] Clifford Nass, Youngme Moon, B.J. Fogg, Byron Reeves, and D.Christopher Dryer. 1995. Can computer personalities be human personalities? *International Journal of Human-Computer Studies* 43, 2 (aug 1995), 223–239. https://doi.org/10.1006/IJHC.1995.1042

[67] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94*. ACM Press, New York, New York, USA, 72–78. https://doi.org/10.1145/191666.191703

[68] Andreea Niculescu, George M. White, See Swee Lan, Ratna Utari Waloejo, and Yoko Kawaguchi. 2008. Impact of English regional accents on user acceptance of voice user interfaces. In *Proceedings of the 5th Nordic conference on Human-computer interaction building bridges - NordiCHI '08*. ACM Press, New York, New York, USA, 523. https://doi.org/10.1145/1463160.1463235

[69] Arika Okrent. 2010. *In the land of invented languages : a celebration of linguistic creativity, madness, and genius*. Spiegel & Grau Trade Paperbacks, New York, USA.

[70] Pierre W. Orelus. 2017. Accentism Exposed: An anticolonial analysis of accent discrimination with some implications for minority languages. In *Language, Race, and Power in Schools: A Critical Discourse Analysis*. Routledge, London, UK, 127–137.

[71] Cathy Pearl. 2016. *Designing Voice User Interfaces : Principles of Conversational Experiences*. O'Reilly Media, Boston, USA.

[72] James Pfrehm. 2018. *Technolingualism : the mind and the machine*. Bloomsbury, Cambridge, Massachusetts. https://www.bloomsbury.com/uk/technolingualism-9781472578365/

[73] Janet B. Pierrehumbert, Tessa Bent, Benjamin Munson, Ann R. Bradlow, and J. Michael Bailey. 2004. The influence of sexual orientation on vowel production (L). *The Journal of the Acoustical Society of America* 116, 4 (oct 2004), 1905–1908. https://doi.org/10.1121/1.1788729

[74] Ian Pitt and Alistair Edwards. 2003. *Design of Speech-based Devices*. Springer London, London, UK. https://doi.org/10.1007/978-1-4471-0093-5

[75] Robert J. Podesva. 2007. Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics* 11, 4 (sep 2007), 478–504. https://doi.org/10.1111/j.1467-9841.2007.00334.x

[76] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, New York, New York, USA, 1–12. https://doi.org/10.1145/3173574.3174214

[77] Eva Reinares-Lara, Josefa D. Martín-Santana, and Clara Muela-Molina. 2016. The Effects of Accent, Differentiation, and Stigmatization on Spokesperson Credibility in Radio Advertising. *Journal of Global Marketing* 29, 1 (jan 2016), 15–28. https://doi.org/10.1080/08911762.2015.1119919

[78] Henry Rogers, Ron Syth, and Greg Jacobs. 2000. Vowel and sibilant duration in gay- and straight- sounding male speech. *International Gender and Language Association 1* (2000).

[79] Debbie A. Rowe. 2010. From Wood to Bits to Silicon Chips: A History of Developments in Computer Synthesized Speech. In *Computer Synthesized Speech Technologies: Tools for Aiding Impairment: Tools for Aiding Impairment*, John Mullennix and Steven Stern (Eds.). IGI Global, New York, USA, 9–27.

[80] Anara Sandygulova and Gregory M.P. O'Hare. 2015. Children's Responses to Genuine Child Synthesized Speech in Child-Robot Interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts - HRI'15 Extended Abstracts*. ACM Press, New York, New York, USA, 81–82. https://doi.org/10.1145/2701973.2702058

[81] Marie Louise Juul Søndergaard and Lone Koefoed Hansen. 2018. Intimate Futures. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18*. ACM Press, New York, New York, USA, 869–880. https://doi.org/10.1145/3196709.3196766

[82] Aaron Springer and Henriette Cramer. 2018. "Play PRBLMS": Identifying and Correcting Less Accessible Content in Voice Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 296, 13 pages. https://doi.org/10.1145/3173574.3173870

[83] Jane Stuart-Smith. 2007. Empirical evidence for gendered speech production: /s/ in Glaswegian. In *Laboratory Phonology 9*, Jennifer Cole and José Ignacio Hualde (Eds.). De Gruyter, New York, USA, 65–86.

[84] Rie Tamagawa, Catherine I. Watson, I. Han Kuo, Bruce A. MacDonald, and Elizabeth Broadbent. 2011. The Effects of Synthesized Voice Accents on User Perceptions of Robots. *International Journal of Social Robotics* 3, 3 (aug 2011), 253–262. https://doi.org/10.1007/s12369-011-0100-4

[85] Andrew R Timming. 2017. The effect of foreign accent on employability: a study of the aural dimensions of aesthetic labour in customer-facing and non-customer-facing jobs. *Work, Employment and Society* 31, 3 (jun 2017), 409–428. https://doi.org/10.1177/0950017016630260

[86] James Tompkinson. 2015. Accent evaluation and the perception of spoken threats. In *Proceedings of the third Postgraduate and Academic Researchers in Linguistics at York conference (PARLAY 2015)*. University of York, UK, 115–131. https://yorkpapersinlinguistics.files.wordpress.com/2016/06/james-tompkinson-parlay-proceedings-2015.pdf

[87] TomTom. 2018. Navigation Voices. https://www.tomtom.com/en_gb/sat-nav/maps-services/shop/navigation-voices/

[88] Rachelle Waksler. 2001. Pitch range and women's sexual orientation. *WORD* 52, 1 (apr 2001), 69–77. https://doi.org/10.1080/00437956.2001.11432508

[89] Paul Warren. 2016. *Uptalk*. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9781316403570

[90] John C. Wells. 1982. *Accents of English 1: An Introduction*. Cambridge University Press, Cambridge, UK.

[91] Nick Wingfield. 2017. 'Cortana, Open Alexa,' Amazon Says. And Microsoft Agrees. https://www.nytimes.com/2017/08/30/technology/amazon-alexa-microsoft-cortana.html

[92] Jessica Wormald. 2016. *Regional Variation in Punjabi-English*. Ph.D. Dissertation. Universty of York, York, UK. http://etheses.whiterose.ac.uk/13188/

[93] Peter Wright and John McCarthy. 2010. *Experience-Centered Design: Designers, Users, and Communities in Dialogue*. Morgan and Claypool Publishers.

[94] Dong Yu and Li Deng. 2015. *Automatic Speech Recognition*. Springer, London, UK. https://doi.org/10.1007/978-1-4471-5779-3

[95] Qing Zhang. 2005. A Chinese yuppie in Beijing: Phonological variation and the construction of a new professional identity. *Language in Society* 34, 3 (2005), 431–466. https://doi.org/10.1017/S0047404505050153